CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models

Year: 2024 | *Citations: 199* | *Authors: Zhihao Du, Yuxuan Wang, Qian Chen*

Abstract

In our previous work, we introduced CosyVoice, a multilingual speech synthesis model based on supervised discrete speech tokens. By employing progressive semantic decoding with two popular generative models, language models (LMs) and Flow Matching, CosyVoice demonstrated high prosody naturalness, content consistency, and speaker similarity in speech in-context learning. Recently, significant progress has been made in multi-modal large language models (LLMs), where the response latency and real-time factor of speech synthesis play a crucial role in the interactive experience. Therefore, in this report, we present an improved streaming speech synthesis model, CosyVoice 2, which incorporates comprehensive and systematic optimizations. Specifically, we introduce finite-scalar quantization to improve the codebook utilization of speech tokens. For the text-speech LM, we streamline the model architecture to allow direct use of a pre-trained LLM as the backbone. In addition, we develop a chunk-aware causal flow matching model to support various synthesis scenarios, enabling both streaming and non-streaming synthesis within a single model. By training on a large-scale multilingual dataset, CosyVoice 2 achieves human-parity naturalness, minimal response latency, and virtually lossless synthesis quality in the streaming mode. We invite readers to listen to the demos at https://funaudiollm.github.io/cosyvoice2.